Community Embeddings for Friend Suggestions

Anton I. N. Begehr
Graduate School of Business
National Research University Higher School of Economics
Moscow, Russia
a.begehr@fu-berlin.de

Prof. Dr. Petr Panfilov
Graduate School of Business
National Research University Higher School of Economics
Moscow, Russia
ppanfilov@hse.ru

***Abstract*—Graphs, such as social networks, emerge naturally from various real-world situations. Recently, graph embedding methods have gained traction in data science research. Recommender systems are used in a wide range of business applications and are essential for online, e-business models to survive and strive in the contemporary market. Using graph embeddings for recommendation tasks, have the possibility of improving upon recommender systems, because of data compression, their feature vector format, and sub-quadratic time complexity. Graph and community embeddings generated with ComE BGMM+VI are used to build a recommender system for friend suggestions. ComE BGMM+VI is an alteration of the community embeddings algorithm ComE. ComE BGMM+VI applies a Bayesian Gaussian mixture model and variational inference for community embedding and detection. Recommendations are evaluated by the top-N hit-rate over users with at least 50 friends. A friend suggestions recommender system with a top-10 leave-one-out hit-rate of 43.6% and run-time optimized 32.9% is presented.**

***Index Terms*—graph, embedding, community embedding, ComE, recommendations, friend suggestions**

I. Introduction

Graphs, such as social networks, knowledge graphs, content-rating graphs, and communication networks, emerge naturally from various real-world situations. Analyzing these graphs leads to findings and understanding of the underlying structures, coherences, and dependencies. Recently, methods for embedding graph's nodes into lower-dimensional Euclidean spaces, called graph embeddings, have gained traction in multiple areas of data science research [6].

Due to the rapid growth of the internet and data accumulation, recommender systems are essential for e-business and online business models to survive and strive in the contemporary market [18]. Modern recommender systems need to take into account the huge amounts of user data generated at all times in big data systems around the world and improve recommendations instead of failing under the thrust of big data overload.

Utilizing graph embeddings for recommendation tasks, has recently gained research traction [16, 17, 7, 20]. The advantages of graph embeddings include data compression and the Euclidean feature vector format [5]. Given these advantages and provided competitive results, graph embeddings have the possibility of greatly improving upon graph-based use-cases like recommender systems.

Community Embeddings, in addition to embedding a graph's nodes through first- and second-order proximity, also preserve higher-order proximity by embedding clusters present in the graph data. The graph and community embedding algorithm ComE aims to preserve first-, second- and higher-order proximity by embedding a graph's nodes and communities [3].

This work specifically examines community embeddings for friend suggestion recommender systems and evaluates recommendations on social network graph data for the use-case of friend suggestions. Graph and community embeddings generated with ComE BGMM+VI are used to develop a friend suggestions recommender system based on the shortest distances between nodes in the embedding. Recommendations are evaluated by the top-N recommendations hit-rate of test edges. A friend suggestions recommender system with a top-10 leave-one-out hit-rate of 43.6% and run-time optimized 32.9% is presented.

II. Friend Suggestions

Recommender Systems are eagerly researched in academia and widely deployed in real-world business applications. Most contemporary technology companies heavily rely on recommender systems to drive usage of their services and consumption of their content. Users, in turn, rely on recommender systems to find what they want and need and save searching time. State-of-the-art recommender systems provide a competitive advantage desperately needed by online services. Companies heavily relying on recommender systems include YouTube, Amazon, Netflix, and many more [19].

Recommender systems are built on top of user-item interaction. Friend suggestions are a simpler type of recommender system. For friend suggestions, no distinction is made between users and items. The entity user is both the subject and object of recommendation. This results in a simple data model, which can be used for structural friend suggestions, as shown in Fig. 1:

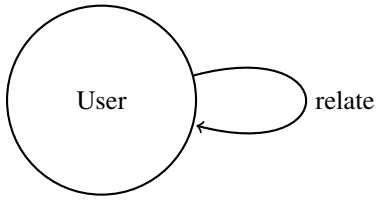

Fig. 1: Minimal data model supporting friend suggestions.

The two main methods used for recommender systems are collaborative filtering and content-based filtering. Collaborative filtering is based on the assumption that users like items similar to other items they like and items that are liked by other users with similar tastes [8, 21]. Content-based filtering considers user and item attributes, instead of solely interactions [19, 12].

In this paper, a method of generating friend suggestions based solely on graph data structure, and not on the node or edge attributes, is presented. The graph of users and friendships is embedded to a lower-dimensional space with the community embedding algorithm ComE BGMM+VI [3, 2, 1]. The approach can be considered a variant of collaborative filtering with the euclidean distance of user embeddings and user community membership as measures for user similarity. The proposed advantage of using such an approach is that community embeddings optimize first-, second-, and higher-order proximity between users in the node embedding.

A. Evaluation

To determine the effectivity of friend suggestions, the generated friend suggestions must be properly validated. In this paper, the top-N approach to evaluating recommender systems with hit-rate as the evaluation metric is chosen.

A user's top-N recommendations is a list of N items to be recommended to a specific user. To evaluate recommender systems with the top-N and hit-rate approach, initially, the dataset is split into train and test data. For each testing user, one relation is left out, according to the leave-one-out method and the model is trained on the remaining training dataset. Once the model is trained, a list of top-N recommendations is generated for each testing user. If the item corresponding to the user is in the user's top-N list, a hit is counted, otherwise, a miss is counted. The hit-rate is defined as the total number of hits divided by the number of testing users.

Utilizing the top-N approach with hit-rate to evaluate recommender systems is advantageous to evaluating recommender systems by a link prediction approach since the top-N approach is more realistic in comparison to actual recommender system use-cases. When you open Netflix, Amazon, YouTube, or Facebook friend suggestions, one or multiple top-N recommendation lists are generated and displayed. If you click on an item and buy, watch, or befriend, that is considered a hit, otherwise, a miss.

The hit-rate metric on top-N recommendations provides a realistic option of evaluating recommender systems [4, 15, 22].

III. Algorithm

The proposed algorithm for generating friend suggestions using community embeddings is detailed and evaluated on time complexity. The initially quadratic runtime of generating friend suggestions for all users is then reduced by a factor of K by utilizing a node's community membership. The two resulting algorithms are presented.

Algorithm 1 describes in pseudocode how recommendations are computed from node embeddings generated by ComE BGMM+VI. The terms friend suggestions and social recommendations are used interchangeably.

Algorithm 1 Top-N Social Recommendations based on Node Embeddings

Require: graph $G = (V, E)$, maximum number of communities K, number of walks γ, walk length ℓ, window size ζ, representation size D, negative context size m, parameters (α, β), number of recommendations N.

Ensure: Top-N recommendations for all nodes R.

1: $\Phi, \Phi', \Pi, (\Psi, \Sigma) \leftarrow ComE(G, K, \gamma, \ell, \zeta, D, m, \alpha, \beta)$
2: **for** $(v, v') \in E$ **do**
3: $\quad F_v \leftarrow F_v \cup \{v'\}$
4: $\quad F_{v'} \leftarrow F_{v'} \cup \{v\}$
5: **end for**
6: **for** $v \in V$ **do**
7: $\quad R_v \leftarrow SortedDict(size = N)$
8: $\quad$ **for** $v' \in V \wedge v' \notin F_v \wedge v' \neq v$ **do**
9: $\quad\quad d \leftarrow \|\phi_v - \phi_{v'}\|$
10: $\quad\quad R_v[d] \leftarrow v'$
11: $\quad$ **end for**
12: $\quad R_v \leftarrow R_v.values()$
13: **end for**

The function $ComE(G, K, \gamma, \ell, \zeta, D, m, \alpha, \beta)$ is the ComE BGMM+VI community embedding algorithm. It returns the node embedding Φ, context embedding Φ', community assignment Π, and community embedding (Ψ, Σ). ComE BGMM+VI is based on the ComE algorithm presented by Cavallari et al. in their paper "Learning Community Embedding with Community Detection and Node Embedding on Graphs" as algorithm 1 on page 381 of the publication's in-proceeding [3]. ComE BGMM+VI takes a Bayesian approach to community embedding by utilizing a Bayesian Gaussian mixture model (BGMM) and variational inference (VI) for community embedding and detection instead of a non-Bayesian Gaussian mixture model (GMM) with expectation maximization (EM) [1].

The runtime of ComE BGMM+VI is equivalent to the runtime of ComE in its big-O notation. ComE's time complexity is $O(|V|\,\gamma\ell + |V| + T_1(T_2|V|\,K + K + |E| + |V|\,\gamma\ell + |V|\,K))$, which is linear in time complexity to the graph size: $O(|V| + |E|)$ (line 1) [3]. All friends of each user are determined in $O(E)$ (lines 2-5). A sorted dictionary of the top-N friend suggestions for each user sorted ascending by distances between the node embeddings is computed in $O(|V|^2\,N))$ (lines 6-13). Top-N friend suggestions are generated for each user: $O(|V|)$. For each user, all users not befriended currently are considered ($O(|V|)$) and inserted into a sorted dictionary of length N ($O(N)$).

This brings the total time complexity of Algorithm 1 to $O(|V|\,\gamma\ell + |V| + T_1(T_2|V|\,K + K + |E| + |V|\,\gamma\ell + |V|\,K) + |E| + |V|^2\,N)$. We consider γ, ℓ, T_1, T_2, and N as constant, therefore the time complexity depends on the graph's size ($|V|$,$|E|$) and the number of communities K: $O(K + |V|\,K + |E| + |V|^2)$.

TABLE I: Algorithm 1 and 2 time complexity.

K constant	Algorithm 1	$O(\|E\| + \|V\|^2)$
	Algorithm 2	$O(\|E\| + \|V\|^2)$
K scaling	Algorithm 1	$O(K + \|V\| K + \|E\| + \|V\|^2)$
	Algorithm 2	$O(K + \|V\| K + \|E\| + \frac{\|V\|^2}{K})$

Algorithm 2 describes in pseudocode how recommendations are computed from ComE BGMM+VI node and community embeddings and community assignments. The advantage of also considers community membership, is that an improvement in time complexity can be obtained, by considering only users in the same community for friend suggestions.

Algorithm 2 Top-N Social Recommendations based on Node and Community Embeddings

Require: graph $G = (V, E)$, maximum number of communities K, number of walks γ, walk length ℓ, window size ζ, representation size D, negative context size m, parameters (α, β), number of recommendations N.

Ensure: Top-N recommendations for all nodes R.

1: $\Phi, \Phi', \Pi, (\Psi, \Sigma) \leftarrow ComE(G, K, \gamma, \ell, \zeta, D, m, \alpha, \beta)$
2: **for** $v \in V$ **do**
3: $\quad k \leftarrow \pi_v$
4: $\quad C_k \leftarrow C_k \cup \{v\}$
5: **end for**
6: **for** $(v, v') \in E$ **do**
7: $\quad F_v \leftarrow F_v \cup \{v'\}$
8: $\quad F_{v'} \leftarrow F_{v'} \cup \{v\}$
9: **end for**
10: **for** $v \in V$ **do**
11: $\quad R_v \leftarrow SortedDict(size = N)$
12: $\quad$ **for** $v' \in C_k \wedge v' \notin F_v \wedge v' \neq v$ **do**
13: $\quad\quad d \leftarrow \|\phi_v - \phi_{v'}\|$
14: $\quad\quad R_v[d] \leftarrow v'$
15: $\quad$ **end for**
16: $\quad R_v \leftarrow R_v.values()$
17: **end for**

Line 1, lines 6-9, and lines 10-17, except for line 12, are the same in Algorithm 2 as in Algorithm 1. All users are filtered into sets C_k, one for each community k in $O(|V|)$ (lines 2-5). Instead of considering the set of all non-friends users of size $|V|$, only users in the same community C_k of size $\frac{|V|}{K}$ are considered (line 12) in $O(\frac{|V|^2 N}{K})$.

This brings the total time complexity of Algorithm 2 to $O(|V|\,\gamma\ell + |V| + T_1(T_2|V|\,K + K + |E| + |V|\,\gamma\ell + |V|\,K) + |V| + |E| + \frac{|V|^2 N}{K})$. Again, we consider γ, ℓ, T_1, T_2, and N as constant, therefore the time complexity depends on the graph's size ($|V|$,$|E|$) and the number of communities K: $O(K + |V|\,K + |E| + \frac{|V|^2}{K})$.

The runtimes for Algorithm 1 and 2 when considering K as constant or scaling w.r.t the graph $G = (V, E)$ are shown in Table I. The number of friend suggestions N is considered constant w.r.t the graph $G = (V, E)$.

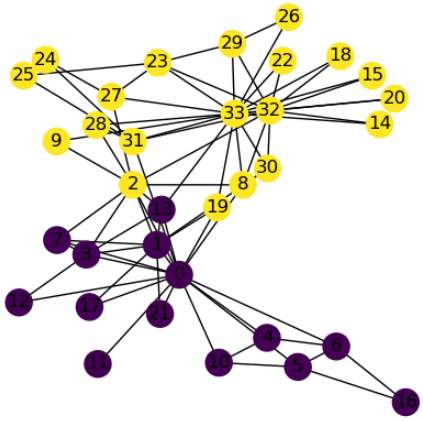

Fig. 2: *Zachary's Karate Club graph* plotted with networkx's spring layout [14, 9].

When the hyperparameter number of communities K is considered non-constant with respect to the graph G, Algorithm 2 reduces the quadratic runtime in comparison to Algorithm 2. K can therefore reduce the quadratic summand $|V|^2 N$ by only considering users of the same community for friend suggestions. The reduction in time complexity through Algorithm 2 is only scalable when considering K to be scale with the input graph's size. This assumption can be made when considering communities as groups of friends: with ten times more users, these users form ten times more communities.

IV. Visual Example

The small *Zachary's Karate Club graph* dataset will be used to visually underline the motivation for using node embedding and community memberships generated by ComE BGMM+VI for generating friend suggestions for person nodes on the karate club graph [14].

Fig. 2 shows the 34 nodes and 78 edges of the *Zachary's Karate Club graph* dataset plotted as a graph with the Python network package networkx's spring layout.[9] Spring layout simulates a force-directed representation of the network, by treating edges as springs pulling nodes together and treating nodes as repelling objects, sometimes called an anti-gravity force.[9] The node classification, represented by node color, shown in Fig. 2 was obtained by running ComE with BGMM and VI with the hyperparameters presented in Table II.

As can be seen in Figure 2, the two communities in the *Zachary's Karate Club graph* dataset are correctly identified, despite the number of communities being initialized with $K = 5$. Thanks to the Bayesian approach to community modeling and optimization through variational inference, three unused communities are dropped, which leaves the prominent two communities present in the dataset. The node classification

TABLE II: Hyperparameters used for the Karateclub dataset.

parameter	notation	value
number_walks	γ	10
walk_length	ℓ	80
representation_size	D	2
num_workers		10
num_iter		3
reg_covar		0.00001
batch_size		50
window_size	ζ	10
negative	m	5
lr		0.025
alpha	α	0.1
beta	β	0.1
down_sampling		0.0
communities	K	5
weight_concentration_prior	Γ	10^{-5}

obtained reflects the node classification published in the 2017 original ComE paper by Cavallari et al.[3]

A. Training and Testing Split

Zachary's Karate Club graph is split into training and testing parts. For evaluation, the leave-one-out (LOO) method is used, where for each testing user one edge is left out, then if the left out edge is in the user's top-N recommendations, a hit is counted, otherwise, a miss is recorded. The leave-one-out method is used in literature for evaluating recommender systems.[22, 10]

The three users with the highest degree are chosen as testing users: user 30, user 0, and user 32. For each testing user, a LOO-testing edge is chosen: (30, 23), (0, 1), and (32, 30). These three edges are spared for the testing dataset and omitted from the training graph,

B. Embedding

ComE BGMM+VI is run with the hyperparameters listed in Table II on the training dataset. The resulting node and community embeddings are visualized as node embeddings colored by community membership and community embeddings plotted as ellipses in Fig. 3.

C. Friend Suggesting

With embeddings generated, top-N friend suggestion recommendations are generated from the training node and community embeddings for the testing users. For each testing user, a ranked list of non-friend users from the same community is generated (Algorithm 2). A test user's friend suggestion list is ordered in ascending order by the euclidean distance between the node embeddings. Table III shows the test users' friend suggestion ranked lists.

The testing users' removed edges make it to the lists successfully; therefore, the removed friends are in the same community as the test users and can be recommended through friend suggestions.

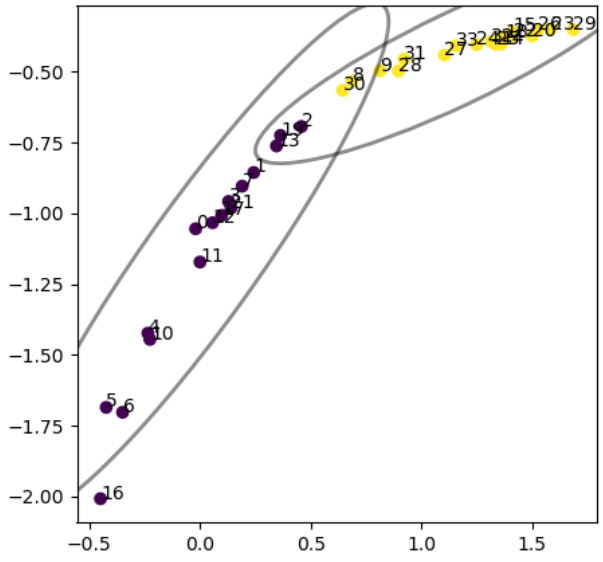

Fig. 3: *Zachary's Karate Club graph* training dataset node and community embedding and node classification.

TABLE III: Testing users' friend suggestions ranked by node embedding distance.

User	Distance
24	0.137181
25	0.214332
23	0.543287

(a) user 33

User	Distance
1	0.322270
16	1.030336

(b) user 0

User	Distance
25	0.039816
24	0.045994
27	0.182484
26	0.255288
28	0.356683
9	0.501317
30	0.668306

(c) user 32

D. Evaluation

The top-N friend suggestions, generated for the testing users with Algorithm 2, are evaluated against the three removed test edges. The ranked lists of friend-suggestions shown in Figure III is used to determine top-N friend suggestions for each testing user and evaluate the recommendations my the hit-rate metric. The LOO-testing edges are printed bold in Table III. The resulting top-N hit-rates are plotted against the number of suggestions N in Fig. 4.

Fig. 4 shows promising results for using ComE to generate

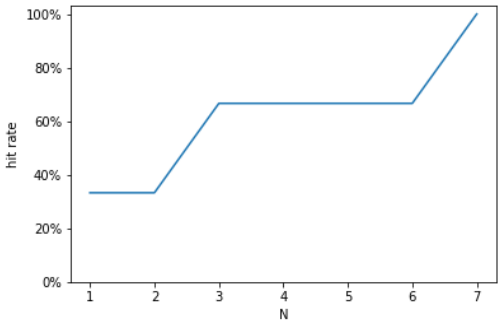

Fig. 4: Hit-rates from top-N recommendations on *Zachary's Karate Club graph*.

top-N friend suggestions on the small *Zachary's Karate Club graph*. The achieved hit-rate starts at 33% for $N = 1$, reaches 67% at $N = 3$, and maxes out at 100% for $N = 7$; therefore, when presenting three friend suggestions to one test user, at least in 66% of cases a friend suggestion would be of interest.

V. Facebook Friend Suggestions

The recommender system based on community embeddings with the Bayesian Gaussian mixture model and variational inference is quantitatively evaluated. Friend suggestions are generated for the *Social circles: Facebook* dataset [11, 13].

The method of generating friend suggestions using ComE BGMM+VI embeddings and the smallest-distance approach is applied to the *Social circles: Facebook* graph. Both Algorithm 1 and Algorithm 2 are evaluated. The top-N recommendations are evaluated by the hit-rate metric.

References

[1] Anton Begehr. *abegehr/ComE_BGMM*. 2020. URL: https://github.com/abegehr/ComE_BGMM (visited on 05/09/2020).

[2] S. Cavallari, E. Cambria, H. Cai, K. C. Chang, and V. W. Zheng. "Embedding Both Finite and Infinite Communities on Graphs [Application Notes]". In: *IEEE Computational Intelligence Magazine* 14.3 (2019), pp. 39–50.

[3] Sandro Cavallari, Vincent W. Zheng, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. "Learning Community Embedding with Community Detection and Node Embedding on Graphs". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM '17. Singapore, Singapore: Association for Computing Machinery, 2017, 377–386. ISBN: 9781450349185. DOI: 10.1145/3132847.3132925. URL: https://doi.org/10.1145/3132847.3132925.

[4] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. "Performance of Recommender Algorithms on Top-N Recommendation Tasks". In: Jan. 2010, pp. 39–46. DOI: 10.1145/1864708.1864721.

[5] Primož Godec. *Graph Embeddings — The Summary*. 2018. URL: https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007 (visited on 08/27/2020).

[6] Palash Goyal and Emilio Ferrara. "Graph Embedding Techniques, Applications, and Performance: A Survey". In: *Knowledge-Based Systems* 151 (July 2018), 78–94. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2018.03.022. URL: http://dx.doi.org/10.1016/j.knosys.2018.03.022.

[7] László Grad-Gyenge, Attila Kiss, and Peter Filzmoser. "Graph Embedding Based Recommendation Techniques on the Knowledge Graph". In: July 2017, pp. 354–359. DOI: 10.1145/3099023.3099096.

[8] Prince Grover. "Various Implementations of Collaborative Filtering". In: *Towards Data Science* (Dec. 28, 2017). URL: https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0 (visited on 08/08/2020).

[9] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11 –15.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. "Neural Collaborative Filtering". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, 173–182. ISBN: 9781450349130. DOI: 10.1145/3038912.3052569. URL: https://doi.org/10.1145/3038912.3052569.

[11] Jure Leskovec. *Social circles: Facebook*. 2012. URL: https://snap.stanford.edu/data/egonets-Facebook.html (visited on 08/15/2020).

[12] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. "Content-based Recommender Systems: State of the Art and Trends". In: Jan. 2011, pp. 73–105. DOI: 10.1007/978-0-387-85820-3_3.

[13] Julian McAuley and Jure Leskovec. "Learning to Discover Social Circles in Ego Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., 2012, 539–547.

[14] Scott Pakin. *Zachary's Karate Club graph*. 1977. URL: https://networkx.github.io/documentation/stable/auto_examples/graph/plot_karate_club.html (visited on 05/09/2020).

[15] Enrico Palumbo, Giuseppe Rizzo, and Raphaël Troncy. "Entity2rec: Learning User-Item Relatedness from Knowledge Graphs for Top-N Item Recommendation". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys '17. New York, NY, USA: Association for Computing Machinery, 2017, 32–36. ISBN: 9781450346528. DOI: 10.1145/3109859.

3109889. URL: https://doi.org/10.1145/3109859.3109889.

[16] Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, Elena Baralis, Michele Osella, and Enrico Ferro. "An Empirical Comparison of Knowledge Graph Embeddings for Item Recommendation". In: *DL4KGS@ESWC*. 2018.

[17] Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, Elena Baralis, Michele Osella, and Enrico Ferro. "Knowledge Graph Embeddings with node2vec for Item Recommendation". In: *ESWC*. 2018.

[18] Nikolaos Polatidis and Christos K. Georgiadis. "Recommender Systems: The Importance of Personalization in E-Business Environments". In: *IJEEI* 4 (2013), pp. 32–46.

[19] Baptiste Rocca. "Introduction to Recommender Systems". In: *Towards Data Science* (June 3, 2019). URL: https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada (visited on 08/08/2020).

[20] Vishwas Sathish, Tanya Mehrotra, Simran Dhinwa, and Bhaskarjyoti Das. "Graph Embedding Based Hybrid Social Recommendation System". In: *ArXiv* abs/1908.09454 (2019).

[21] Xiaoyuan Su and Taghi M. Khoshgoftaar. "A Survey of Collaborative Filtering Techniques". In: *Adv. Artif. Intell.* 2009 (2009), 421425:1–421425:19.

[22] Z. Zhao, Ming Zhu, Y. Sheng, and Jinlin Wang. "A Top-N-Balanced Sequential Recommendation Based on Recurrent Network". In: *IEICE Trans. Inf. Syst.* 102-D (2019), pp. 737–744.